Environmental Health Perspectives Supplements
Vol. 102 Suppl 1:91–94 1994

Some Comments on Potency Measures in Mutagenicity Research

by Barry H. Margolin,[1] Byung Soo Kim,[2] Melissa G. Smith,[1] Bethel A. Fetterman,[1] Walter W. Piegorsch,[3] and Errol Zeiger[4]

In this article, the measurement of the potency of a chemical or mixture from its dose response in a particular assay is addressed. Attention is focused on data from the Ames Salmonella assay. Three measures of potency are explored and shown to be highly correlated. The presentation then discusses specific areas of research that might benefit from a study of potency.

Introduction

For more than a decade, the National Institute of Environmental Health Sciences (NIEHS) has been developing an extensive multitest genetic toxicology database. The most notable feature of this database is a set of results from the application of four commonly used *in vitro* short-term tests (STTs) to 114 chemicals for which National Toxicology Program (NTP) 2-year rodent carcinogenicity assay data are available. The four STTs are mutagenesis in Salmonella (SAL) and mouse lymphoma cells (MLA) and chromosome aberrations (ABS) and sister chromatid exchanges (SCE) in Chinese hamster ovary cells. The first major analyses performed on this database focused on 73 chemicals whose testing for carcinogenicity by the NTP was completed during the period December 1976 to January 1985 (*1*). These analyses focused primarily on the qualitative predictivity of rodent carcinogenicity from the four *in vitro* STTs. The major conclusions of that study were: *a*) Qualitative concordances of the four STTs with rodent carcinogenicity did not show significant differences among assays (all approximately 60%) and were much lower than previous estimates. *b*) A negative STT was not predictive of noncarcinogenicity; a positive SAL, on the other hand, was somewhat predictive of carcinogenicity, but positives in the other three tests were less so. *c*) There was no complementarity among the STTs, and no battery of tests constructed from two or more of these four STTs improved upon the carcinogen predictivity of the SAL test alone.

These conclusions elicited varied reactions within the genetic toxicology community. Some individuals felt there must be something erroneous in the findings; after all, during the previous decade there had been numerous publications reporting concordances of 90% or better for SAL. Two criticisms did appear worthy of further investigation. The first was that the 73 chemicals in the initial investigation were in some way atypical, and therefore replication of the findings was needed from a second set of chemicals. The second criticism was that statistical analyses had primarily focused on the qualitative (positive/negative) results obtained for the 73 chemicals, and that an analysis that included quantitative results, e.g., measures of potency, might lead to different conclusions regarding the predictivity of rodent carcinogenicity from STTs.

The first criticism was effectively answered by the publication of the results of a follow-up study of an additional 41 chemicals the NTP had tested for both rodent carcinogenicity and genetic toxicity using the four STTs listed above (*2,3*). These papers confirmed the major conclusions drawn by Tennant et al. (*1*) regarding the lack of complementarity among the four STTs and the inability of a battery drawn from these four assays to improve upon the Salmonella assay for predicting rodent carcinogenicity. Interestingly, the initial 73-chemical study and the 41-chemical follow-up demonstrated no statistically significant differences between the two data sets on any relevant dimension. Consequently, as it serves the purposes of this paper, the 73-chemical and 41-chemical data sets will be treated either as an initial study and a replicate or as one combined study of 114 chemicals; the former will permit exploratory analyses of the 73-chemical data

[1] Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7400.

[2] Department of Applied Statistics, Yonsei University, Seoul, 120-749, South Korea.

[3] Statistics and Biomathematics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709.

[4] Experimental Toxicology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709.

Address reprint requests to B. H. Margolin, Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7400.

This paper was presented at the International Biostatistics Conference on the Study of Toxicology that was held May 13–25, 1991, in Tokyo, Japan.

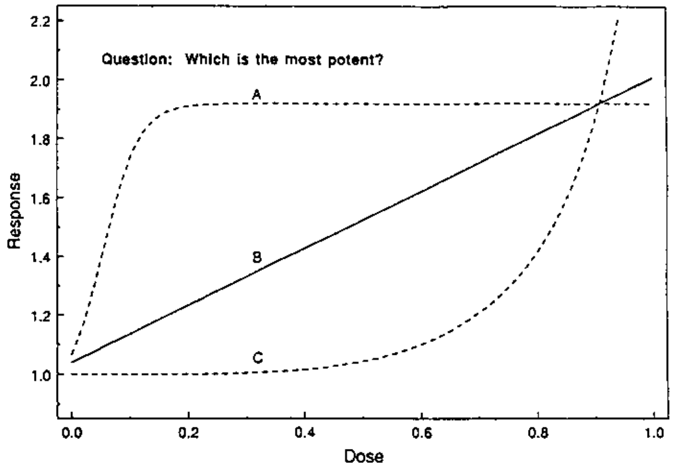

FIGURE 1. Three theoretical dose–response curves, all intersecting at a common high dose. The question: which is the most potent of the three?

set, with the 41-chemical data set reserved for purposes of validation of any important findings.

Measures of Potency

The second criticism discussed above, that use of quantitative information would improve carcinogenicity predictivity of the four *in vitro* STTs, was still unanswered. It focused attention on the need for the development and evaluation of measures of potency for each of the four STTs involved in the NTP studies. On reflection, it is clear that there is no unique and universal way to measure the potency of a chemical in an assay unless all dose–response curves for that assay share the same shape, e.g., linear. If the shapes of dose–response curves vary from chemical to chemical, as they do in the real world, then the selection of a measure of potency to characterize an observed dose–response curve is not a straightforward matter. Consider the three dose–response curves pictured in Figure 1. Which is the most potent of the three? Given that the purpose of this line of research is to address problems of human health, the argument might be made that behavior at low doses, the most common human exposure, should be used for measuring potency. Ideally then, one would want to know the incremental change in, and hence the derivative of, the dose–response curve at low dose. Although this argument is attractive, it may be fallacious for purposes of predicting rodent carcinogenicity. In the work to be described, as well as in ongoing work, various measures of potency are considered.

This paper is an interim report on the development and evaluation of potency measures for each of the four STTs and the interrelations of these measures. Initial efforts have focused on the SAL assay, for which proposals for measuring the potency of the response observed have been published (*4,5*). Three measures of potency have been considered initially:

1. The point-rejection estimate of Bernstein et al. (*4*) is predicated on an assumption of low-dose linearity of the dose response. In the computation of this estimate, observations that depart from the assumed linearity are discarded in turn from the highest dose to the lowest, with the slope recomputed after each discard. The slope of the regression of the remaining observations on dose yields the measure of mutagenic potency, which is labeled b_B. Computer code to evaluate this measure was graciously provided by Bernstein et al. (*4*).
2. Margolin et al. (*5*) describe an estimate of mutagenic potency that is based on a class of nonlinear dose–response models of the Ames assay. The models describe the probability p(D) that a plated bacterium will give rise to a visible revertant colony, given that the plate on which it was placed was exposed to dose D of the test chemical:

$$p(D) = \{1 - \exp[-(\alpha + \beta D)]\} \cdot T(D) \ .$$

Here T(D) is the function describing the toxicity to the bacterium of dose D of the test chemical. The two forms for T(D) considered were:

$$T(D) = \exp(-\gamma D)$$

or

$$T(D) = [2 - \exp(\gamma D)]_+$$

where $[x]_+ = \max(x, 0)$. In this model of the SAL assay response, the parameter β reflects the mutagenic effect per unit dose, adjusted for concomitant toxicity. The estimate of β presented in Margolin et al. (*5*) is denoted by b_M.
3. The maximum observed slope is defined as the maximum average colony count per plate per dose achieved at any of the positive doses in an experiment. If Y_k is the mean number of revertants per plate observed at dose D_k, for $0 = D_0 < D_1 < \ldots < D_r$, for r positive doses, the estimate is given by:

$$\max_k \ [(\overline{Y}_k - \overline{Y}_0)/(D_k - D_0)]$$

This estimate of potency is denoted by b_Y.

The first two potency measures above are specific to the SAL assay because they attempt to estimate low-dose mutagenic potency after adjustment for toxicity. As mentioned, the argument for this focus on low dose is that it more nearly reflects the typical human exposure. The third method of estimating potency is generic and directly applicable to each of the four STTs under study. This report will briefiy discuss the behavior of these three measures of potency for the SAL assay. Three other generic measures of potency applicable to dose–response curves from all assays are also under study but they will not be reported here. These generic measures are:

4. The simple slope obtained by a linear regression of the observed SAL colony count/plate on the dose, ignoring considerations of nonlinearity due to toxicity and other phenomena. This estimate is denoted by b_L.
5. The lowest effective dose (LED), i.e., the lowest dose yielding an effect that is statistically significantly elevated over the control value. In the particular implementation studied, adjustment is made for multiple comparisons. In using this measure, it is tacitly acknowledged that one is a captive of the doses that are used in the experiment, i.e., this

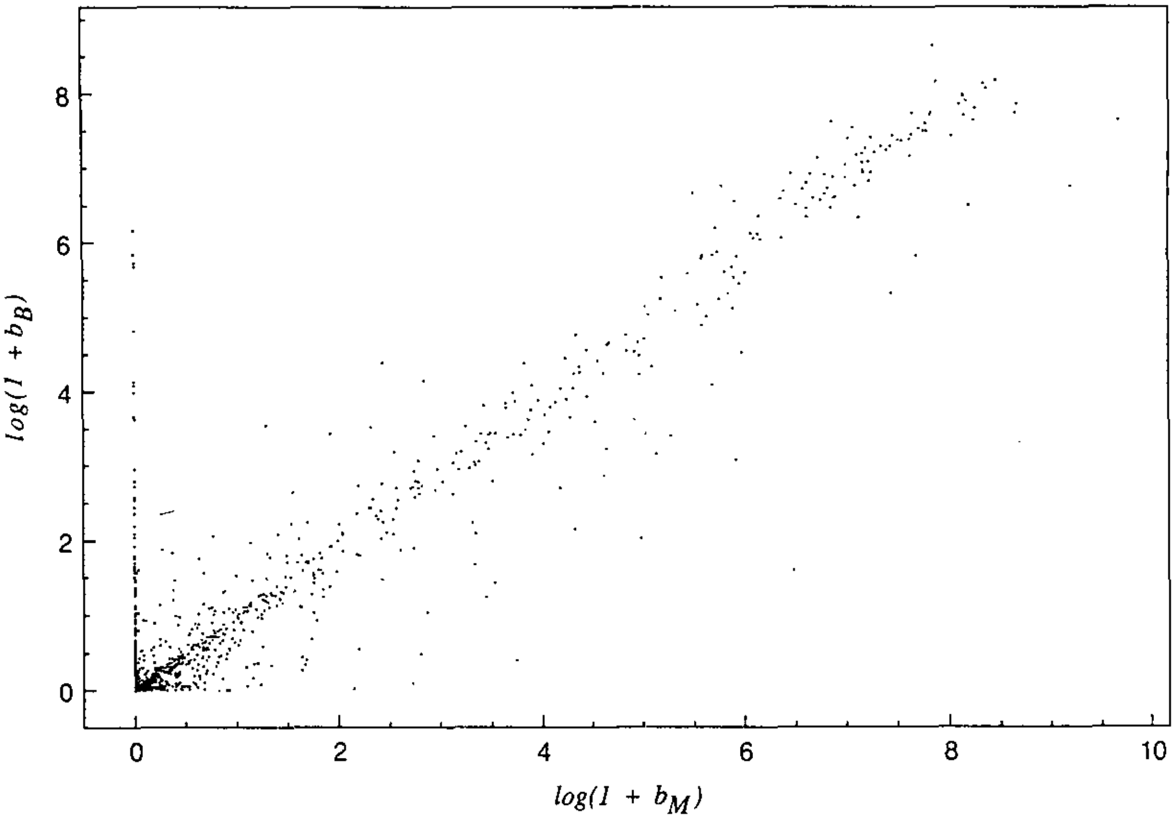

FIGURE 2. Plot of $\log(1+b_B)$ versus $\log(1+b_M)$.

measure can only assume values equal to one of the doses employed.

6. The measure of potency of Margolin and Risko (6), which estimates the dose needed to induce a "unit" increase (DUI) over control for a dose response that is assumed to be a second-degree polynomial. By definition, this measure is arbitrary in its definition of a unit increase. Nevertheless, within any one assay, it produces a credible measure of relative potency.

Descriptive Statistics for the Potency Measures b_B, b_M, and b_X

For the 73-chemical data set, there are 2613 experiments available for evaluation of b_B, b_M, and b_X, whereas for the 41-chemical data set there are 1464. The estimator b_X is always defined, as long as there is one treated dose and a control. This is not the case for b_B or b_M. For the evaluation of b_B, at least two treated doses must remain after exclusion, in turn, of the high doses that departed from linearity of the response. For the estimation of b_M, there must be at least two treated doses available for fitting the nonlinear model after exclusion of high doses that exhibit very substantial toxicity; for elaboration of this point, see Margolin et al. (7). Although it is clear that these last two measures of potency may not always be computable, the results below on percentage computability of the measures for the NTP data are surprising with regard to b_B (Table 1).

Table 1. Percent of experiments for which measure is computable.

Data set	b_B	b_M	b_X
73 chemicals	77	94	100
41 chemicals	77	94	100

Thus, B_B cannot be computed for 23% of either database, a surprisingly high percentage.

For each of the three measures of potency, the distribution of estimated chemical potencies across the database is highly skewed. Logging the three potency measures yields better-behaved random variables that are much less skewed. Further evidence for this claim for logging the measures can be seen in the correlation matrixes for the three measures, with and without logging. For the unlogged measures, the observed correlation matrix is shown in Table 2. The correlation matrix for logged measures is shown in Table 3.

Table 2. Correlation matrix for unlogged measures.

	b_B	b_M	b_X
b_B	1.00	0.74	0.78
b_M		1.00	0.81
b_X			1.00

Table 3. Correlation matrix for logged measures.

	b_B	b_M	b_X
b_B	1.00	0.95	0.94
b_M		1.00	0.90
b_X			1.00

Figure 2 is a scatterplot of log $(1+b_B)$ versus $\log(1+b_M)$ showing the strong linear relationship between the two. This strong correlation is reassuring because both measures strive to estimate the slope of the same dose-response curve at low dose. For situations in which one of these two measures has been used in an evaluation, such as in Piegorsch and Hoel (*8*), one would not expect much change if the other measure were substituted.

Uses of Measures of Potency

The study of measures of potency was initially motivated by the desire to predict rodent carcinogenicity from quantitative measures of mutagenicity, i.e., mutagenic potency of chemicals for individual assays or for a battery of assays. This can and will be studied either with carcinogenicity itself remaining a qualitative variable or with carcinogenicity being treated in a quantitative manner as well. A specific application of the former would be an extension of the Carcinogenicity Prediction Battery Selection (CPBS)* methodology of Rosenkranz et al. (*9*) to include measures of potency of short-term assays. For a discussion of one extension to include dependent, qualitative variables, plus a list of associated references for CPBS, see Kim and Margolin (*10*).

Another possible use of a measure of mutagenic potency occurs in certain epidemiological studies, where urines of individual subjects are tested via the Salmonella assay for signs of exposure to environmental toxicants (*11*). An example would be the study of oncology nurses who are responsible for delivering antineoplastic treatments (*12*). If multiple concentrations of a subject's urine are tested in this assay, a dose-response curve is obtained. A measure of potency from the dose-response curve would facilitate analyses more sophisticated than simply recording whether a response at least two times background was observed, and then proceeding to analyze this dichotomous variable. One could anticipate greater study power deriving from the use of a potency measure, which in turn would permit the use of smaller study sample sizes.

Another use of measures of mutagenic potency will be to assess the potencies observed for the chemicals tested in the ongoing NTP "Sea of Mutagens" study. In this study, a representative sample of 100 chemicals has been drawn from the approximately 50,000 synthetic chemicals introduced to commerce in the last 45 years. Although it is not feasible to test in a timely fashion each of the 100 chemicals in a 2-year rodent carcinogenicity assay, each can be tested for mutagenicity using Salmonella. Results from this study will address the question of whether the human race is awash in a sea of mutagens of its own creation. The use of a measure of potency in this NTP study will permit a refinement of the objective in which not all mutagens will be treated equally. If a certain percentage of the 100 chemicals is found to be positive, it will be most informative to know the distribution of the potencies for the positives.

Finally, studying the distribution of potencies in the NTP database will shed light on the reason for the observation among NTP toxicologists that in certain STTs, positive results are not always reproducible. If one assumes that the analysis of an STT does not have an inherently elevated false positive level, the major factor controlling the reproducibility of a positive response is the power of the assay to detect an effect. This power will vary from chemical to chemical. With the use of a measure of potency for a particular STT, one can formulate an approximate distribution of power for the chemicals in one's database and then proceed to investigate the probability of reproducing an initial positive response for the chemicals in question. Intuitively, strongly acting chemicals will have high power and a high level of reproducibility, whereas weakly acting agents will have low power and low reproducibility. The use of a measure of potency will enable one to quantify in probabilistic terms this important issue of test result reproducibility.

Portions of the work of Barry H. Margolin, Byung Soo Kim, Melissa G. Smith, and Bethel A. Fetterman were supported by contract no. 293-90-1-0005 from the National Institute of Environmental Health Sciences. Melissa G. Smith had additional partial support from a training grant in environmental biostatistics (no. 5 T32 ES07018), funded by the National Institute of Environmental Health Sciences; and Bethel A. Fetteman received additional support from a National Science Foundation Graduate Fellowship.

REFERENCES

1. Tennant, R. W., Margolin, B. H., Shelby, M. D., Zeiger, E., Haseman, J. K., Spalding, J., Caspary, W., Resnick, M., Stasiewicz, S., Anderson, B, and Minor, R. Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. Science 236: 933-941 (1987).
2. Haseman, J. K., Zeiger, E., Shelby, M. D., Margolin, B. H., and Tennant, R. W. Predicting rodent carcinogenicity from four in vitro genetic toxicity assays: An evaluation of 114 chemicals studied by the National Toxicology Program. J. Am. Stat. Assoc. 85: 964-971 (1990).
3. Zeiger, E., Haseman, J. K., Shelby, M. D., Margolin, B. H., and Tennant, R. W. Evaluation of four in vitro genetic toxicity tests for predicting rodent carcinogenicity: confirmation of earlier results with 41 additional chemicals. Environ. Mol. Mutagen. 16 (suppl. 18): 1-14 (1990).
4. Bernstein, L., Kaldor, J., McCann, J., and Pike, M. C. An empirical approach to the statistical analysis of mutagenesis data from the Salmonella test. Mutat. Res. 97: 267-281 (1982).
5. Margolin, B. H., Kaplan, N., and Zeiger, E. Statistical analysis of the Ames Salmonella/microsome test. Proc. Natl. Acad. Sci. U.S.A. 78: 3779-3783 (1981).
6. Margolin, B. H., and Risko, K. J. The statistical analysis of in vivo genotoxicity data: case studies of the rat hepatocyte UDS and mouse bone marrow micronucleus assays. In: Evaluation of Short-Term Tests for Carcinogenicity: Report of the International Program on Chemical Safety Collaborative Study on In Vivo Assays, Vol. 1 (J. Ashby, F. J. deSerres, M. D. Shelby, B. H. Margolin, M. Ishidate, Jr., and G. C. Becking, Eds.). Cambridge University Press, Cambridge, 1988, chapter 3, pp. 29-42.
7. Margolin, B. H., Kim, B. S., and Risko, K. J. The Ames Salmonella/microsome mutagenicity assay: issues of inference and validation. J. Am. Stat. Assoc. 84: 651-661 (1989).
8. Piegorsch, W. W., and Hoel, D. G. Exploring relationships between mutagenic and carcinogenic potencies. Mutat. Res. 196: 161-175 (1988).
9. Rosenkranz, H. S., Klopman, G., Chankong, V., Pet-Edwards, J., and Haimes, Y. Y. Prediction of environmental carcinogens: a strategy for the mid-1980s. Environ. Mutagen. 6: 231-258 (1984).
10. Kim, B. S., and Margolin, B. H. The prediction of carcinogenicity by batteries of dependent short-term tests. Environ. Health Perspect. 101: 000-000 (1992).
11. Margolin, B. H. Statistical aspects of using biologic markers. Stat. Sci. 3: 351-357 (1988).
12. Rogers, B., and Emmett, E. A. Handling antineoplastic agents: urine mutagenicity in nurses. IMAGE: J. Nursing Schol. 19: 108-113 (1987).

*CPBS is a registered trademark of Case Western Reserve University.